Finding and evaluating sets of nearest neighbours

Julie Weeds and David Weir
School of Cognitive and Computing Sciences
University of Sussex, Brighton, BN1 9QH, UK

Abstract

In this paper, we consider two applications of distributional similarity measures, probability estimation and prediction of semantic similarity. We investigate whether high performance in one application area is correlated with high performance in the other. This work also provides an evaluation of two state-of-the-art distributional similarity measures and introduces a variant of one. Further, we overcome statistical biases in the standard pseudo-disambiguation task and look at the effect of word and co-occurrence frequency on the performance of the measures.

1. Introduction

Two words are said to be distributionally similar if they appear in similar contexts. However, such a loose definition leads to as many questions as it answers. For example, we can ask what constitutes context and whether all contexts should be treated equally. Accordingly, many techniques have been developed which use corpora to determine sets of distributionally similar words. These range from standard geometric measures, such as the L_1 Norm, to information-theoretic measures, such as the one proposed by Lin (1998).

Since the loose definition of distributional similarity also leads to a problem in evaluating the measures, evaluation is often based on the ability of sets of nearest neighbours, determined using the measures under consideration, to perform a given application-based task (Lin 1998; Lee 1999). However, there are many potential applications for sets of distributionally similar words and it is not clear whether the same properties which make a measure useful in one application will make it useful in another. One distinction that can be made between applications is whether they make use of the property of distributional similarity directly or as a predictor of semantic similarity. For example, in a probability estimation task (e.g. Dagan et al. 1999), where the aim is to predict the probabilities of unseen word co-occurrences, the property of distributional similarity can be used directly. Several researchers have proposed a semantic class based approach to the PP-attachment ambiguity resolution problem (e.g. Resnik 1993). However, this is in fact a syntactic task which might benefit from the words in classes being syntactically or distributionally related rather than semantically related. On the other hand, distributional similarity is also proposed as a predictor of semantic similarity and has therefore been used in tasks such as automatic thesaurus generation (e.g. Lin 1998) and word-for-word glossing (Pantel and Lin 2000). Further, as a predictor of semantic similarity, there may be potential for using sets of distributionally similar words in tasks which currently rely on a hand-crafted thesaurus such as WordNet (Fellbaum 1998) e.g. malapropism correction (Budanitsky and Hirst 2001), text simplification (Carroll et al. 1998), collocation extraction (Pearce 2002) and text summarization (Silber and McCoy 2002).

We investigate whether high performance on a pseudo-disambiguation task, a purely distributional task often used as an evaluation task for probability estimation (Dagan et al. 1999), is correlated with high semantic similarity. In order to do this, we compare the performance of three distributional similarity measures and one semantic similarity measure on a pseudo-disambiguation task and also the ability of the distributional similarity measures to predict semantic similarity, as defined by the semantic similarity measure.

Two measures which are widely used (e.g. Schulte Im Walde and Brew 2002; Wiebe 2000) and have been shown to outperform other similarity measures are Lin's mutual information (MI) based measure (Lin 1998) and Lee's α-skew divergence measure (Lee 1999). We also present a variation of Lin's MI-based measure which we will theoretically justify and will show performs slightly better than the standard MI-based measure on the pseudo-disambiguation task. However, not only do we consider which measure is best for the given application, we also seek to establish what properties of the measure leads to the better performance. In particular, we look at the interaction of a word's frequency and the quality of its nearest neighbours.

The rest of this paper is organised as follows. First, the four similarity measures used throughout our experiments will be introduced. We will then discuss obtaining sets of neighbours from the data extracted from the BNC for two disjoint sets of nouns. This is followed by a discussion of the pseudo-

disambiguation task and results on this task for all four measures. We will then compare the sets of neighbours derived using the distributional similarity measures with those derived using the semantic similarity measure. Finally, the effect of low frequency events on all three distributional similarity measures will be considered.

2. Similarity measures

In this section, each of the four similarity measures used throughout this work will be introduced in turn. However, it should be noted that when using any of the three distributional similarity measures, the description (or distribution) of a word is taken to be a set of dependency relation co-occurrences and frequencies. Further, we consider only similarity between nouns and describe nouns in terms of the verbs with which they co-occur in the direct-object relation. Accordingly, we use the notation $P(v|n)$ to refer to the probability of the verb in a direct-object co-occurrence pair being v given that the noun is n. However, it would be possible, with all three measures, to consider other relations, other parts of speech and even other types of context.

2.1. The α-skew divergence measure (Lee, 1999)

Given two nouns, n_1 and n_2, and the notation $q(v) = P(v|n_1)$ and $r(v) = P(v|n_2)$ where v ranges over all verbs, the distance between the two nouns can be computed as the distance between the two probability distributions as defined by the α-skew divergence measure, $sim_{asd}(r,q)$.

$$sim_{asd}(r,q) = D(q \,||\, \alpha.r + (1 - \alpha).q)$$

where

$$D(q \,||\, r) = \sum_v q(v) \log \frac{q(v)}{r(v)}$$

The value of the alpha-skew divergence measure is zero if the two probability distributions are identical and increases positively as the distributions become less similar. The parameter α ranges between 0 and 1 and controls the extent to which the measure approximates D (the Kullback-Leibler divergence measure). When α is close to 1, the measure closely approximates the Kullback-Leibler divergence measure whilst avoiding the problem of zero probabilities associated with using the Kullback-Leibler divergence measure in NLP. Accordingly, we use α=0.99. Further, the measure is asymmetric and it defines how well the probability distribution associated with n_2 approximates the probability distribution associated with n_1. Consequently it can be used to determine how fit n_2 is as a neighbour for n_1.

2.2. The MI-based measure (Lin, 1998)

The MI-based measure is based on Lin's theorem that the similarity between two objects is defined to be the amount of information in the *features* they have in common divided by the sum of the information in the *features* of each object. If the only grammatical relation being considered is the direct-object relation *dobj*, n is a noun and v is a verb, v can be considered to be a feature of the noun n if $I(n,v)$, the mutual information between n and v, is positive. The mutual information between two words occurring in the direct-object relation can be computed using the formula (Hindle 1990):

$$I(n,v) = \log \frac{P(v,n)}{P(v)P(n)} = \log \frac{P(n \mid v)}{P(n)}$$

In other words, v is a feature of n if our expectation of seeing n increases on seeing v. Accordingly, if $T(n)$ is the set of verbs v that are features of n, the similarity between the two nouns, n_1 and n_2 can be calculated as:

$$sim_{mi}(n_1,n_2) = \frac{\sum_{v \in T(n_1) \cap T(n_2)} (I(n_1,v) + I(n_2,v))}{\sum_{v \in T(n_1)} I(n_1,v) + \sum_{v \in T(n_2)} I(n_2,v)}$$

The value of the MI-based measure lies in the range [0,1] where a value of 1 indicates that the two

nouns are identical (Lin 1997).

2.3. Variation on the MI-based Measure

Lin's MI-based measure considers two nouns to be identical if they have exactly the same features. It is possible for two nouns to have different probability distributions yet be considered identical. This is because two nouns are considered to completely share a feature regardless of the extent to which each noun has that feature (i.e. the value of the mutual information between the noun and the feature verb). The difference in the extent to which each noun has a particular shared feature can be considered so that only nouns with identical probability distributions are considered to be identical. To do this, the similarity between two nouns, n_1 and n_2, can be defined as:

$$sim_{mi_{diff}}(n_1, n_2) = \frac{\sum_{v \in T(n_1) \cap T(n_2)} (I(n_1, v) + I(n_2, v) - |I(n_1, v) - I(n_2, v)|)}{\sum_{v \in T(n_1)} I(n_1, v) + \sum_{v \in T(n_2)} I(n_2, v)}$$

The expression on the numerator of the equation can be rewritten in the following way:

$$A + B - |A - B| = 2 \times \min(A, B)$$

and thus, this variation has parallels with the measure defined in Hindle (1990):

$$sim_{Hindle} = \sum_{v \in T(n_1) \cap T(n_2)} \min(I(n_1, v), I(n_2, v))$$

2.4. A WordNet Based Measure (Lin, 1997)

There are many different measures of semantic similarity (see Budanitsky (1999) for a thorough review). Here, we use a measure which is based on the hyponymy relation (as defined by WordNet), proposed by Lin (1997), which has been shown to perform fairly well compared to other WordNet based measures on a spelling correction task (Budanitsky and Hirst 2001).

If *S(n)* is the set of senses of the noun *n* in WordNet, *sup(c)* is the set of (possibly indirect) superclasses of concept *c* in WordNet then the similarity between n_1 and n_2 can be calculated as:

$$sim_{wn}(n_1, n_2) = \max_{c_1 \in S(n_1), c_2 \in S(n_2)} \left[\max_{c \in \sup(c_1) \cap \sup(c_2)} \frac{2 \log P(c)}{\log(P(c_1)) + \log(P(c_2))} \right]$$

In other words, the *commonality* (Lin 1997) of two concepts (or synsets) is defined as the maximally specific common superclass of those concepts. The similarity of two nouns is then defined as the similarity between their most similar senses. *P(c)* is the probability that a randomly selected noun refers to an instance of *c*. These probabilities *P(c)* are estimated by the frequency of concepts in SemCor (Miller et al. 1994), a sense-tagged subset of the Brown corpus, noting that the occurrence of a concept refers to instances of all the superclasses of that concept.

3. Finding Sets of Neighbours

Much previous evaluation of similarity measures has involved determining neighbours for the most frequently occurring nouns in a particular corpus (Lin 1998; Lee 1999). However, since we wish to investigate the effects of word frequency on similarity, we determine neighbours of 1000 high frequency nouns and also of 1000 low frequency nouns.

2,852,300 lemmatised (noun,verb) direct-object pairs were extracted from the BNC using a parser developed by Briscoe and Carroll (1995, 1996). Having discarded all co-occurrence pairs where the noun does not occur in WordNet, we constructed two sets denoted: $nouns_{highfreq}$ made up of the pairs where the noun is one of the 1000 most frequently occurring nouns, and $nouns_{lowfreq}$ made up of the pairs where the noun occurs between 70 and 120 times in the data set[1].

For each noun, we randomly select 80% of the available data as training data and set aside the other 20% as test data for the pseudo-disambiguation task. We compute the similarity between each pair of nouns using each of the four similarity measures under consideration. For each noun and similarity measure, a set of ranked neighbours is constructed.

[1] These frequencies correspond to ranks of 3001 to 4000.

4. Pseudo-Disambiguation Experiment

Pseudo-disambiguation experiments have become a standard method of evaluating a technique's ability to perform probability estimation (e.g. Pereira et al. 1993, Lee 1999 and Clark and Weir 2002). In the current context, we may use a word's neighbours to decide which of two co-occurrences (one of which occurred in the test data and the other of which did not) is more likely. However, various approaches can be taken to the way training/test sets are constructed and the way in which neighbours vote on their preferred co-occurrence and this will now be discussed.

4.1. Test Set Construction

In this section, we will first describe how the test set for our experiments was constructed and then explain the rationale behind it. Having set aside 20% of the data for each noun as test data, we converted each noun-verb (n,v_1) pair in that data into a noun-verb-verb (n,v_1,v_2) triple where $P(v_1)$ is approximately equal to $P(v_2)$ over all of the training data and (n,v_2) has not been seen in the test or training data. We selected ten test instances[2] for each noun as follows. Whilst more than ten triples remained, duplicate triples were discarded. If there were still more than ten triples, after all duplicates had been discarded, ten triples were selected at random from the remaining triples. This process resulted in two sets of 10000 test triples. Each set of test triples was then split into five disjoint sets (containing 2 triples for each noun) so that five-fold cross validation could be performed.

We now consider approaches taken by earlier researchers. In order to ensure that the test data is unseen, both Pereira et al. (1993) and Clark and Weir (2002) randomly select a portion of the co-occurrences as test data and then delete any instances of these co-occurrences from the training data. However, we feel that this approach is inappropriate for the large-scale evaluation of distributional similarity measures since completely removing co-occurrence types from the training data distorts the data used by the measures to determine similarity. A different approach, taken by Lee (1999), is to randomly select a portion of the co-occurrences as test data and then delete any instances of co-occurrences from this test data which also occur in the training data. The next step, in either approach, is to replace each remaining noun-verb (n,v_1) pair in the test data with a noun-verb-verb triple (n,v_1,v_2) where v_2 is selected such that it is of approximately equal probability to v_1 in the training data. It is then standard to split test sets into five disjoint sets for cross-validation purposes. Neighbours of n are then required to decide which was the original co-occurrence which, by construction, is always (n,v_1).

Although discarding test pairs which appear in the training data ensures that the test data is unseen and overcomes the problem of removing items of training data, we argue that there are remaining problems to be considered.

First, there is no guarantee that the inequality $P(v_1|n)>P(v_2|n)$ holds for population probabilities (as opposed to sample probabilities) even if we introduce the restriction that (n,v_2) is unseen in the training data. For (n,v_1) not to have been seen in the training data (and thus remain in the test data), our estimate of the probability of this co-occurrence must be very small and is therefore not much greater than that of the completely unseen, in either training or test data, (n,v_2). Accordingly, if we were to see a lot more data, (n,v_2) may become more probable/frequent than (n,v_1), and, even if it does not, we are still expecting neighbours to distinguish between events of very similar probabilities.

Second, since the joint probability of n and v_1 must be small (for the co-occurrence pair not to have appeared in the training data), the unigram probability of the verbs to be disambiguated tends to decrease as the unigram probability (or frequency) of the noun increases[3]. This in turn limits the number of neighbours the verbs may have occurred with and in turn may influence the performance results.

Third, the test set, and therefore the average performance, is biased towards nouns which occur with many different verbs (typically high frequency nouns). In the extreme case, after discarding those pairs which also occur in the training data, there may be no pairs remaining for a very selective noun. Accordingly, we are not evaluating over the whole set of nouns. When 1000 random nouns were selected and a test set constructed in the way described by Lee (1999), we found that only 60% of the nouns appeared in the final test set.

Our method of test set construction overcomes all three of these problems. We argue that there is no need to discard any test data on the basis of it having been seen in the training data (or vice versa). In the machine-learning paradigm, it is still unseen data which is not used in the training process. Further, by not discarding seen noun-verb pairs we can be more confident that the correct choice verb

[2] Ten being less than the minimum number (fourteen) of triples in the test data for any noun.

[3] In our experiments, we found significant negative correlation (Pearson's product-moment correlation coefficient between -0.12 and –0.18) between noun frequency and verb frequency.

(v_1) is more plausible than the incorrect choice verb (v_2) and we do not introduce any statistical biases into the test set i.e. there is no significant negative correlation between noun and verb frequency. Further, we select a proportion of the data for each noun as test data which means that each noun is treated fairly as to how much data is lost for training purposes whilst allowing us to control the minimum number of test instances for each noun. Accordingly, we were able to remove the bias towards less selective/high frequency nouns.

4.2. Nearest Neighbour Voting

As mentioned in the previous section, the task is for the nearest neighbours to decide which of (n,v_1) and (n,v_2) was the original co-occurrence. In both Pereira et al. (1993) and Lee (1999), the k-nearest neighbours, where k is a parameter which is optimised, of the noun n vote as to which is the most likely co-occurrence (each neighbour m votes for the co-occurrence according to which verb it co-occurs with most in the training data i.e. by considering whether $P(v_1|m)>P(v_2|m)$). Performance is then measured as error rate.

$$error\ rate = \frac{1}{T}\left(\#of\ incorrect\ choices + \frac{(\#of\ ties)}{2}\right)$$

where T is the number of test instances and a tie results when the neighbours cannot decide between alternatives.

However, in earlier work (Weeds 2002), we recognised that giving each neighbour a single vote does not allow us to distinguish between cases where a neighbour occurs with each verb approximately the same number of times and where a neighbour occurs with one verb significantly more times than the other. We also observed that a single vote per neighbour appears to introduce a bias towards v_1. When all neighbours are considered, and therefore the two verbs have appeared approximately the same number of times, v_1 tends to win more votes by virtue of the fact that it has occurred with more different nouns (whereas v_2 wins its votes by a larger amount). Thus, we give each neighbour m a vote which is equal to the difference in frequencies of the co-occurrences (m,v_1) and (m,v_2). Performance is still measured as error rate.

4.3. Results

Table 1 shows the optimal error rates for each similarity measure when used to determine neighbours for high and low frequency nouns. This optimal error rate is given in terms of the mean error rate at the optimal value of k averaged over the five test sets used for the five-fold cross validation. The standard deviations of these means, not given, were of the order of 0.01.

Measure	**Noun Frequency**			
	high		**low**	
	k	**opt. error rate**	***k***	**opt. error rate**
asd	30	0.230	60	0.192
mi	50	0.193	100	0.181
mi$_{diff}$	40	0.192	130	0.176
wn	20	0.295	40	0.294

Table 1

We make the following observations based on Table 1.

1. All of the distributional similarity measures significantly outperform the WordNet based measure on the pseudo-disambiguation task. This is regardless of the frequency of the original noun.
2. Both MI-based measures significantly outperform the α-skew divergence measure on this task.
3. The use of the mi_{diff} measure leads to a small improvement over the standard MI-based measure for low frequency nouns.
4. All of the distributional similarity measures perform best for low frequency nouns whereas there is no significant difference when using the WordNet based measure.

We will return to these observations later.

5. Semantic Similarity

The results in the previous section suggest that semantic similarity is not necessarily the best indicator of neighbours' abilities to perform probability estimation, which could be considered a purely syntactic

task. However, many potential applications of nearest neighbour sets in NLP are in the semantic domain and therefore would require words to be semantically similar. We evaluated how well each distributional similarity measure predicts semantic similarity by measuring the similarity between neighbour sets generated using the distributional similarity measures and neighbour sets generated using the WordNet based measure.

In order to compare two neighbour sets, we convert each neighbour set so that each neighbour is given a score based on its rank order. We do not use the similarity directly in our comparison since 1) this would require us to convert the α-skew divergence measure distances into similarities and 2) many tasks, including the earlier pseudo-disambiguation task, make no use of the actual similarity scores. Neighbours are given a rank score of *(k+1)-rank* where *k* is the number of neighbours being considered and *rank* is the neighbour's rank in the neighbour set[4]. Having performed this transformation, we use the same calculation as Lin (1998) (except for simplifications due to the use of ranks) to compute the similarity between two neighbour sets. Supposing two neighbour sets for the same word, *w*, are represented by two ordered sets of words $[w_k....w_1]$ and $[w_k'....w_1']$, then their similarity is defined as:

$$\frac{\sum_{w_i = w_j', i \in j} i}{\sum_{i=1}^{k} i^2}$$

This is similar to calculating a rank order correlation coefficient except for the fact that it does not require the same words to be in both sets. We could in fact calculate a rank order correlation coefficient based on the entire neighbour sets (i.e. where k=2000) [5] but we cannot do so directly for smaller subsets. Our method of scoring, however, means that we are approximating a product-moment coefficient between the ranks where all neighbours not within a given rank distance of the noun tie on the rank zero.

In our experiments, the *k* nearest neighbours of each noun, computed using each of the distributional similarity measures, were compared with the *k* nearest neighbours of the noun according to the WordNet based measure (with no restrictions on the frequencies of the neighbours). The mean similarity between neighbour sets for k=200 for both the high frequency nouns and for the low frequency nouns are given in Table 2. The standard deviations of the means, not given, were all of the order of 0.1. k=200 was chosen in order to consider more than just closest neighbours whilst avoiding going into the tail of dissimilar words (which may be ranked in any order). However, similar results were obtained in experiments considering 50 and 100 nearest neighbours.

Measure	**Noun Frequency**	
	high	**low**
asd	0.290	0.270
mi	0.307	0.210
mi_{diff}	0.303	0.172

Table 2

From these results we make the following observations:

1. In terms of predicting semantic similarity, the MI-based measures outperform the α-skew divergence measure for high frequency nouns.
2. However, the α-skew divergence measure performs significantly better than the MI-based measures for low frequency nouns.
3. Similarity with the WordNet based measure is, in all cases, greater for high frequency nouns than low frequency nouns.

Returning to the observations made in Section 4, it is interesting to note the differences with respect to high and low frequency nouns. In particular, semantic similarity is in all cases lower for low frequency nouns whilst performance on the pseudo-disambiguation task is higher for low frequency nouns. Further, there is a significant drop in performance of the MI-based measures for low frequency nouns in terms of semantic similarity.

6. Dependence on Low Frequency Events

We now explore why the MI-based measures perform poorly at the semantic evaluation for low

[4] Using a reverse-order rank score means that all words outside the neighbour set can be given the score 0

[5] we do not, however, wish to calculate the correlation over all neighbours since, generally, we are not concerned about the order in which a measure ranks dissimilar words.

frequency nouns. We predict that it is because low frequency nouns are made too similar to each other. This would have less effect on the pseudo-disambiguation task since low frequency neighbours will have occurred with fewer verbs than their higher frequency counterparts. Further, we hypothesise that the unduly high similarity between low frequency nouns is because of the large influence of low frequency co-occurrences on the MI-based measures. We investigated the first of these predictions by comparing sets of neighbours where the neighbours are restricted in frequency and the second using a Bayesian Estimation technique.

6.1. Neighbour Frequency

For each noun and similarity measure, two new sets of ranked neighbours were constructed; one set where neighbour nouns are restricted to being in the $nouns_{highfreq}$ set and one set where neighbour nouns are restricted to being in the $nouns_{lowfreq}$ set. We then used the neighbour set comparison technique described in Section 5 to compare the similarity between neighbour sets for a single measure. By comparing the complete neighbour set (where neighbours are of any frequency) to the restricted frequency neighbour sets, we get an estimate of how the complete neighbour set is made up of high and low frequency nouns. The results, in terms of mean similarity of the complete neighbour set to the high frequency neighbour set and to the low frequency set, for each of the four measures, are given in Table 3. Standard deviations of the means, not given, were all to the order of 0.1.

Measure	**Noun Frequency**			
	high		**low**	
	sim high	**sim low**	**sim high**	**sim low**
asd	0.992	0.021	0.956	0.102
mi	0.931	0.161	0.459	0.748
mi_{diff}	0.950	0.122	0.209	0.915
wn	0.845	0.330	0.798	0.392

Table 3

From the results in Table 3, we make the following observations:

1. From the results for the WordNet based measure, it can be seen that both high and low frequency nouns tend to be more semantically similar to high frequency nouns than low frequency nouns.
2. There is also a very strong tendency for all of the distributional similarity measures to select high frequency nouns as neighbours of high frequency nouns.
3. The same tendency is seen for the alpha-skew divergence measure for low frequency nouns but it is reversed for the MI-based measures i.e. these measures have a greater tendency to select low frequency nouns as neighbours of low frequency nouns.

This reversal in observation 3 explains the poor performance of the MI-based measures for low frequency nouns on the semantic evaluation task. We will now investigate why the MI-based measures tend to select low frequency neighbours for low frequency nouns.

6.2. Bayesian Estimation

Our second hypothesis was that the MI-based measures cause low frequency nouns to be overly similar to each other due to the large influence of low frequency co-occurrences of low frequency words.

Let us consider the impact of a singular co-occurrence (i.e. a co-occurrence which is seen only once in the corpus) on mutual information and hence the similarity between nouns. From the definition of mutual information, a singular co-occurrence of a low frequency noun and a low frequency verb will lead to higher mutual information between that noun and that verb than if either noun or verb had been of a higher frequency. However, singular co-occurrences are a fairly unreliable source of information. It could be that that co-occurrence should not have occurred (it was due to an error in the corpus) or that it has a much lower probability than (1*/size of corpus)* but, because of discretization in the data, we either see such low probability events once or zero times in the corpus. As a direct consequence, events with the same population probabilities may be assigned either very high mutual information scores or negative mutual information scores. Further, if two low frequency nouns both have a singular co-occurrence with the same low frequency verb, this will lead to high similarity between those nouns.

The dependence of such measures on low frequency events has been considered before. Lee (2001), based on a suggestion made in Katz (1987), discards all singular co-occurrences from the training data.

Discarding all of these co-occurrences leads to a fairly small reduction in data set size in terms of indistinct co-occurrences (12% for our data) but a much larger reduction in terms of distinct co-occurrence types. 66% of co-occurrence types (i.e. distinct co-occurrences), which appear in our BNC co-occurrence data, occur only once. Over the 2000 nouns considered in our experiments, 57% of co-occurrence types occur only once. Further, the description of a low frequency noun may not contain many non-singular co-occurrences – so if we discard all singular co-occurrences, we will be left with the task of trying to determine similarities between low frequency nouns based on their co-occurrences with such frequently occurring verbs as *have* and *make*. It is fairly obvious, that these singular co-occurrences have to have some influence when we are considering low frequency nouns – however, it is our hypothesis that they have too much influence.

The approach taken here is to tackle the root of the problem – that is that we do not have the true population probabilities of the co-occurrences or the words, only corpus-based maximum likelihood estimates. We use a Bayesian Estimation technique to smooth the sample probability distributions.

6.3. A Uniform Prior

Bayesian Estimation allows us to combine a prior distribution (what was expected before any data was observed) with an observed distribution to form a posterior distribution. In the simplest case, where the prior distribution is a uniform distribution, Bayesian Estimation reduces to add1 smoothing (see e.g. Friedman 1999). In our case, a uniform distribution has been assumed since, before any data is seen, we have no idea which noun-verb co-occurrences are more likely than others. However, rather than adding 1 to every possible noun-verb combination, we add a small fraction X so that we can control the total amount of probability mass which is assigned to unseen events. For our data, we use X=0.011 since adding this amount to every possible noun-verb combination is consistent with a total probability mass of 0.118 being assigned to previously unseen events. This value of 0.118 is P_o as calculated using simple Good-Turing techniques (Gale and Sampson 1996) and is based on the number of hapax legomena in the data.

Accordingly, having added 0.011 to every noun-verb combination, we recalculated the similarities between each pair of nouns using each distributional similarity measure. We then compared the neighbour sets generated for each measure (as in Section 6.1).

6.4. Results

Measure	Noun Frequency			
	high		low	
	sim high	sim low	sim high	sim low
asd	0.992	0.022	0.981	0.050
mi	0.953	0.081	0	1
mi_{diff}	0.9996	0.001	0	1

Table 4

Comparing the results in Table 4 with those in Table 3, we can observe that all of the tendencies seen in Table 3 have been exacerbated. However, the tendency of the MI-based measures to select low frequency neighbours for low frequency nouns has been increased by far the most significantly. In fact, the MI-based measures now select neighbours for low frequency nouns exclusively from the low frequency noun set. Accordingly, we can conclude that the Bayesian estimation has had most effect on the MI-based measures for low frequency nouns.

The effects of the Bayesian smoothing on the similarity measures can also be seen by examining the results for the new neighbour sets on the pseudo-disambiguation task (Table 5) and the semantic evaluation against WordNet (Table 6).

Measure	Noun Frequency			
	high		low	
	k	opt. error rate	k	opt. error rate
asd	30	0.228	40	0.199
mi	20	0.261	300	0.313
mi_{diff}	20	0.266	700	0.389

Table 5

Measure	Noun Frequency	
	high	low
asd	0.292	0.268
mi	0.228	0.058
mi_{diff}	0.226	0.047

Table 6

By comparing the results in Tables 5 and 6 with their earlier counterparts (Tables 1 and 2), we can see again that the Bayesian Estimation has had a much greater (negative) effect on the performance of the MI-based measures than the α-skew divergence measure.

6.5. Discussion

In order to be able to understand the effects seen in Section 6.4, we need to consider carefully what is happening when we perform the Bayesian Estimation. Rather than reducing the impact of low frequency events, as might intuitively be expected, we have in fact created a large number of very low frequency events. The results for a measure which relies heavily on low frequency events will therefore be more affected than those for a measure which does not. Even though the frequency of co-occurrence added for unseen events (0.011) is very small, it is considered significant by the MI-based measures particularly when the nouns and verbs under consideration are low frequency ones. Let us consider an example. Suppose *n* and *v* have each occurred 60 times in the training data[6] which contains 1596798 (indistinct) noun-verb co-occurrences but zero co-occurrences of *n* and *v*. There are a total of 2000 nouns in the data and 9587 verbs, so after adding 0.011 to every noun-verb combination, our frequency estimates are:

$f'(n_1,v_1) = 0.011$
$f'(n_1) = 60 + 0.011 \times 9587 = 165.5$
$f'(v_1) = 60 + 0.011 \times 2000 = 82$
$\sum_{v,n} f'(v,n) = 1596798 + 0.011 \times (2000 \times 9587) = 1807712$

The estimated mutual information between n_1 and v_1 (see Section 2.2) is 0.551. Thus, a low frequency verb which has not appeared with our low frequency noun has become a feature of that noun. Further, we can see that all low frequency verbs will become features of all low frequency nouns and hence low frequency nouns will tend to become more similar to each other. Further, in order for new features not to be created by the smoothing process, it was calculated that the parameter X would have to be very small (less than 6×10^{-7} if we allow for nouns and verbs only occurring once). In other words, it would have to be so small that it would have no effect on the mutual information scores.

7. Conclusions and Further Work

All of the distributional similarity measures perform the pseudo-disambiguation task better than the semantic similarity measure. Further, Lin's MI-based measure performs better than the α-skew divergence measure and the variant of Lin's MI-based measure presented here leads to further small improvements for low frequency nouns.

One might have expected that the α-skew divergence measure would therefore perform the best at the semantic evaluation. However, this is not the case. The MI-based measures' neighbour sets for high frequency words are closer to the WordNet derived neighbour sets than the α-skew divergence measure's neighbour sets are.

The MI-based measures do, however, perform poorly at the semantic evaluation for low frequency nouns. We have explored this thoroughly and conclude that it is because low frequency nouns are made overly similar to each other due to the large influence of low frequency co-occurrences of low frequency events on the MI-based measures. This effect is a lot less for the α-skew divergence measure, which is why its neighbours of low frequency nouns are closer in terms of semantic relatedness.

Our final observation is that the frequency of a neighbour influences how much weight it has in the pseudo-disambiguation task. Low frequency neighbours will have occurred with fewer verbs and will therefore contribute to fewer decisions. One consequence of this is that more neighbours will generally need to be considered in order to reach a decision. A second consequence is that, in order to obtain high performance on the pseudo-disambiguation task, we find we can sacrifice some semantic similarity in favour of finding lower frequency potential neighbours, safe in the knowledge that these neighbours, if wrong, will affect less test instance outcomes.

In the future, we plan to look at ways of reducing the effects of word frequency on the distributional similarity measures. For example, Bayesian estimation with a non-uniform or hierarchical prior could be used and thus we could eliminate the problem of introducing extra features into the MI-based measures. We also intend to look at other applications of distributional similarity measures and how the frequency effects observed affect these applications.

[6] which corresponds to *n* occurring 75 times in the complete data set (training + test data)

8. Acknowledgements

We would like to thank John Carroll for the use of his parser and acknowledge that this work is supported by an Engineering and Physical Sciences Research Council studentship to the first author.

References

Briscoe E, Carroll J 1995 Developing and evaluating a probabilistic lr parser of part-of-speech and punctuation labels. In *Proceedings of 4th ACL/SIGDAT International Workshop on Parsing Technologies*, pp 48—58

Budanitsky A, Hirst G 2001 Semantic distance in WordNet: an experimental, application-oriented evaluation of five measures. In *Proceedings of NAACL-01*

Budanitsky A 1999 *Lexical Semantic Relatedness and its Application in Natural Language Processing.* Unpublished PhD thesis, University of Toronto

Carroll J, Briscoe E 1996 Apportioning development effort in a probabilistic lr parsing system through evaluation. In *Proceedings of ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing,* pp 92—100

Carroll J, Minnen G, Canning Y, Devlin S, Tait J 1998 Practical simplification of English text to assist aphasic readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*

Clark S, Weir D 2002 Class-based probability estimation using a semantic hierarchy. *Computational Linguistics* 28(2)

Dagan I, Lee L, Pereira F 1999 Similarity-based models of word co-occurrence probabilities. *Machine Learning* 34(1-3)

Fellbaum C (ed) 1998 *WordNet: an electronic lexical database*. Cambridge, Massachusetts, MIT Press

Friedman N, Singer Y 1999 Efficient Bayesian parameter estimation in large discrete domains. *Advances in Neural Information Processing Systems 11.*

Gale W, Sampson G 1996 Good-Turing estimation without tears. *COGS Research Paper,* Brighton

Hindle D 1990 Noun classification from predicate-argument structures. In *Proceedings of ACL-90,* pp 268—275

Katz S 1987 Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing,* ASSP-35(3), pp 400—401

Lee L 1999 Measures of distributional similarity. In *Proceedings of ACL-99*

Lee L 2001 On the effectiveness of the skew divergence for statistical language analysis. *Artificial Intelligence and Statistics*:pp 65—72

Lin D 1997 Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97.* pp 64—71

Lin D 1998 Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL'98*

Miller G, Chodorow M, Landes S, Leacock C, Thomas R 1994 Using a semantic concordance for sense identification. In *Proceedings of ARPA Human Language Technology Workshop*

Pantel P, Lin D 2000 Word-for-word glossing of contextually similar words. In *Proceedings of ANLP-NAACL'00.* pp 78—85

Pearce D 2002 A comparative evaluation of collocation extraction techniques. In *Proceedings of LREC'02*

Pereira F, Tishby N, Lee L 1993 Distributional clustering of similar words. In *Proceedings of ACL-93*

Resnik P 1993 *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania

Schulte Im Walde S, Brew C 2002 Inducing German semantic verb classes from purely syntactic subcategorization information. In *Proceedings of ACL-02.*

Silber H, McCoy, K 2002 Efficiently computed lexical chains as an intermediate representation for automatic text summarization. In *Computational Linguistics* 28(4)

Weeds J 2002 The reliability of a similarity measure. In *Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK5)*

Wiebe J 2000 Learning subjective adjectives from corpora. In *Proceedings of AAAI'00.*